R E P O R T   R E S U M E S

ED 010 993 SE 000 027

COOPERATIVE DEVELOPMENT OF LOCALLY ORIENTED ACHIEVEMENT TESTS IN CHEMISTRY.

BY- SCHMITT, J.A. AND OTHERS

PUB DATE 66

EDRS PRICE MF-$0.09 HC-$0.44 11P.

DESCRIPTORS- *ACHIEVEMENT TESTS, *CHEMISTRY, *COGNITIVE DEVELOPMENT, *STUDENT EVALUATION, *SECONDARY SCHOOL SCIENCE, EVALUATION, TEST CONSTRUCTION, CHEMISTRY REGENTS EXAMINATION, NEW YORK

HIGH SCHOOL CHEMISTRY UNIT TESTS AND A PRETEST DESIGNED TO PREDICT STUDENT SUCCESS ON A STATE CHEMISTRY EXAMINATION WERE CONSTRUCTED. HIGH SCHOOL SCIENCE TEACHERS AND UNIVERSITY SCIENCE EDUCATORS USED BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES TO CLASSIFY TEST QUESTIONS ACCORDING TO THE DEGREE THAT THEY INVOLVED RECALL, COMPREHENSION, APPLICATION, OR HIGHER COMPETENCIES. RESULTS FROM TRIAL USE OF THE INSTRUMENTS WERE ANALYZED THROUGH ITEM ANALYSIS. REVISED FORMS OF THE TEST WERE USED AS REGULAR EXAMINATIONS FOR APPROXIMATELY 1,200 STUDENTS DURING 1 SCHOOL YEAR. CORRELATIONS BETWEEN UNIT TESTS, AND UNIT TESTS AND PRETESTS WERE DETERMINED. MULTIVARIATE ANALYSIS WAS USED TO IDENTIFY COMBINATIONS OF TESTS THAT WERE THE BEST PREDICTORS OF STUDENT SUCCESS ON THE STATE EXAMINATION. CORRELATIONS WERE ALSO DETERMINED FOR STUDENTS' TOTAL SCORES ON DIFFERENT CLASSES OF QUESTIONS ON ALL TESTS. INTERCORRELATIONS BETWEEN THE UNIT TESTS AND THE PRETESTS RANGED FROM .45 TO .78. THOSE FOR THE UNIT TESTS RANGED FROM .61 TO .78. MULTIPLE CORRELATION COEFFICIENTS FOR DIFFERENT COMBINATIONS OF PREDICTING INSTRUMENTS RANGED FROM .75 TO .84. INTERCORRELATIONS FOR THE FOUR CATEGORIES OF TEST QUESTIONS RANGED FROM .80 TO .87. THIS ARTICLE IS PUBLISHED IN THE "JOURNAL OF RESEARCH IN SCIENCE TEACHING," VOLUME 4, ISSUE 2, 1966. (AG)

ED010993

JOURNAL OF RESEARCH IN SCIENCE TEACHING

Volume 4 — Issue 2 — 1966

CONTENTS

Published under the auspices of
NATIONAL ASSOCIATION FOR RESEARCH IN SCIENCE TEACHING *and*
ASSOCIATION FOR THE EDUCATION OF TEACHERS IN SCIENCE *by*
John Wiley & Sons, Inc., New York • London • Sydney

JOURNAL OF RESEARCH IN SCIENCE TEACHING

Published quarterly (March, June, September, December) by John Wiley & Sons, Inc. covering one volume annually. Publication Office at 20th and Northampton Streets, Easton, Pennsylvania 18043. Executive, Editorial, and Circulation Offices at 605 Third Avenue, New York, New York 10016. Application for Second-class postage pending at Easton, Pa. Subscription price, $10.00 per volume. Single issue price, $3.00. Foreign postage, $0.50 per volume. Manuscripts should be submitted in duplicate to Prof. J. Stanley Marshall, Department of Science Education, The Florida State University, Tallahassee, Florida 32306.

27

JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 4, PP. 85-91 (1966)

The authors describe a cooperative effort to produce reliable tests with high content validity for the schools involved.

Cooperative Development of Locally Oriented Achievement Tests in Chemistry

J. A. SCHMITT
Boston College, Chestnut Hill, Massachusetts

J. J. MONTEAN
The University of Rochester, Rochester, New York

S. S. WINTER and S. D. FARR
State University of New York at Buffalo, New York

Thorndike and Hagen,[1] Ahmann and Glock,[2] Nunnally,[3] and others have noted the limitations of commercially published achievement tests when they are used in determining the extent to which curricular goals peculiar to a given school, school district, or geographic region have been achieved. The dilemma of the test publisher, who must produce a subject matter test general enough to cover diverse curricular goals but specific enough to achieve individual sales, has its counterpart with the teacher or supervisor, who must choose between a well-constructed and field-tested instrument with limited content validity for his particular course, and the technically less sound instrument that he or his department might develop to meet specific local needs.

There are unit tests in chemistry available commercially, of course, but with the exception of one series designed for use with advanced-level college students,[4] these instruments are intended to be used in connection with a particular textbook. No one instrument among them, therefore, could be appropriate for different schools using different texts, and considering the structure of the New York State Regents Syllabus, it seems unlikely that any single series could adequately represent the required content of high school chemistry, as most teachers of that subject in New York view it. Since the standard Regents Examination in any course is an important consideration to teachers throughout the State, moreover, the practice of constructing unit tests for local use, using old Regents tests, is quite common; it is expected that practice with Regents-type items will enhance pupils' prospects of success with the final examination.

Two major limitations face the teacher who is sincerely interested in developing a series of sound, subject matter tests with specifications consistent with his own particular course objectives. The writing of good test items, first of all, has been properly called an art" ... demanding the utmost degree of creativity, ingenuity, and persistence on the part of anyone who practices it,"[5] and to these requirements—most practitioners would allow—should be added a sizable quantity of time; few individuals in teaching will find themselves, simultaneously, in possession of all four.

The project described here attempted to overcome these limitations by bringing together a group of teachers to provide a pool of time, creativity, ingenuity, and persistence with a group of professors from two neighboring universities, to provide technical assistance. Twelve teachers of high school chemistry and four university professors participated throughout the two years of the study (1961–1963); three additional chemistry teachers joined the effort during the second year. Preliminary field testing of the

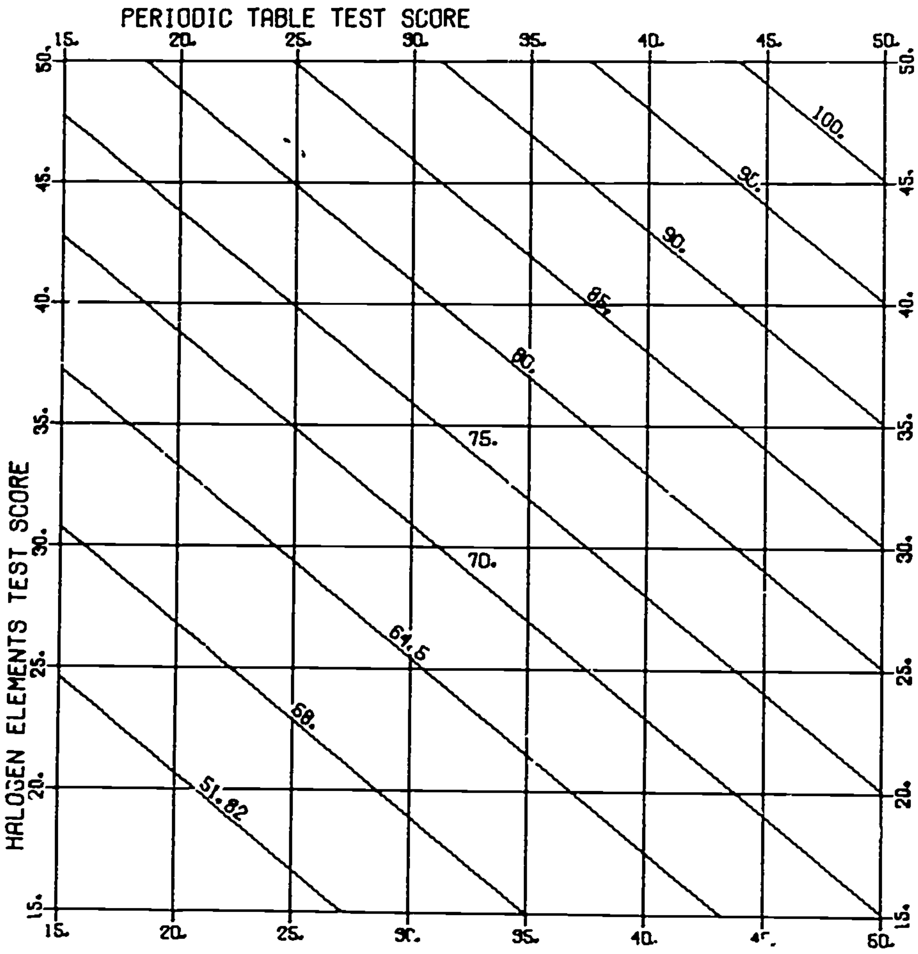

Fig. 1. Regression equation with halogen elements test score plotted against periodic table test score. 0.79 = periodic table test score; 1 = halogen elements test score; 15.38 = constant = chemistry grade (+ OR − 9.75).

seven tests was accomplished during the 1963–1964 school year.

Objective

The objective of the project was to produce an instrument that would contribute significantly to the prediction of success in high-school chemistry courses following the syllabus approved by the Regents of the University of the State of New York, and to develop a series of unit tests that would measure accomplishment with respect to selected areas of the Regents Syllabus. The unit tests were intended to provide checkpoint data for the evaluation of individual and group progress, and to permit analysis of the contributions made to total accomplishment by higher order mental processes, as well as by recall. Both the public school and the university personnel participating in the study were agreed that the Regents Examination in Chemistry (the standard final examination for pupils enrolled in courses following the Regents Syllabus) placed more emphasis on the recall process than could be justified in terms of such frequently mentioned goals of chemistry instruction as development of the abilities to reason scientifically, to think critically, and to solve problems efficiently.

Bloom and his co-workers had already recognized this need, of course, and had attempted to contribute to its fulfillment through their *Taxonomy of Educational Objectives*.[6] So far as measurement in science education is concerned, however, the only use to which the writers of this paper could find the *Taxonomy* having been put—at least in the first five years following its publication—was in connection with the compilation of a folio of test items.[7] Viewing the *Taxonomy* as the best available guide

for the measurement of cognitive competencies other than recall, the group agreed at its first meeting to attempt to incorporate test items representative of Bloom's categories in each of the unit tests.

Procedures in Developing the Pretest and Unit Tests

At the initial meeting of the group, six units of study from the Regents Syllabus were agreed upon as progress checkpoints that would be appropriate in terms of the teaching patterns followed by all of the participating teachers. The teachers then submitted to the consultants tests and test items—mostly from their files of old tests—related to these selected study units, to provide item pools for the tests. Concurrently, and independently, each teacher responded to a questionnaire prepared by the consultants, indicating the relative emphasis given to discrete topics within each of the selected study units. Responses to the questionnaire were averaged to provide the tables of specifications for the unit tests, the teachers having agreed in advance to this procedure.

The test items supplied by the teachers were edited by the consultants and were converted, where necessary, to multiple choice questions. Approximately one hundred items were prepared for each of the selected study units, with the content distribution following, as closely as possible, that of the appropriate table of specifications. Particular care was taken to include items which could be rated according to the several cognitive competencies of the *Taxonomy*.

It was necessary to make several compromises in adapting the *Taxonomy* for use in this project. The considerable detail of the *Taxonomy* could not be fully exploited in unit tests which were to be capable of administration within a single, 45-minute class session, and so it was decided that only the major toxonomical subdivisions should be employed. Secondly, it was found (to no one's surprise) that items requiring analysis, synthesis, or evaluation are quite difficult to construct in the multiple choice form that had been agreed upon. The categorical definitions, finally, are relatively cumbersome and a somewhat simpler classification scheme was decided upon. Accordingly, the following four categories were established for the purposes of the project:

Recall: Any item which had been taught in substantially the same form as that in which it appeared in the test, requiring mere resurrection of a particular bit of information. Rephrasing, inversion of sentences, and similar form changes do *not* remove an item from the recall category.

Comprehension: Any item requiring the application of a principle under circumstances different from those constituting the teaching context of the principle, but in such form that the correct principle is implied in the question.

Application: Similar to comprehension, but the required principle is not implied in the question, so that the student must select the appropriate principle from his repertoire of learned principles, as well as apply it correctly. Quantitative problems were considered in this category.

Higher Competencies: This category included items which required analysis of a complex situation and the subsequent drawing of analytic, synthetic, or evaluative inferences.

Individual test items were categorized according to the above classification scheme by the science education consultants, and these independent judgments were found to be identical with respect to a substantial majority of the items. In cases of disagreement, an attempt was made to reconcile differences of opinion through conference, and failing this, the item in question was discarded. While the writers felt that the degree of judgmental correspondence achieved was adequate to sustain the operational adequacy of the system, they also concluded that the taxonomical categories other than recall, at least as they were employed in this study, are not without ambiguity. Differentiation of recall-type items from

TABLE I
Distribution of Item Types among the Unit Tests

Unit Test	Item type: Recall	Comprehension	Application	Higher order
Periodic table	15	12	10	13
Water, solutions, and ionization	17	15	12	6
The halogen elements	15	18	8	9
Sulfur and nitrogen	22	10	8	10
Principles of organic chemistry	14	30	5	1
Principles of chemical reactions	12	14	18	6
Totals	95	99	61	45

other types, on the other hand, proved to be a simple matter.

Two preliminary forms of each unit test were then constructed and distributed to the participating teachers for class administration at a time considered to be appropriate by the teacher in terms of individual class progress. Each of the preliminary forms consisted of fifty items, and upon the completion of all administrations of the tests pertaining to a particular unit, item analyses were performed to determine item difficulty, discriminating efficiency, and the presence of noneffective foils. At a general meeting of the cooperating group following the completion of preliminary trial of each unit test, test items which had proved to be confusing to pupils or to present other similar problems were eliminated from consideration. Items finally selected for inclusion in each of the unit tests had difficulty indices ranging from 0.35 to 0.80 (proportion passing), had foils of demonstrated effectiveness, and discriminated significantly on the basis of total test score. In its final form, each unit test consisted of fifty items, distributed according to item type as shown in Table I.

In developing the chemistry pretest, the primary goal was production of an instrument that would permit efficient forecasting of Regents Examination results. In consideration of this goal, content analysis of immediate past Regents Exams was employed to construct a table of specifications, and items were almost entirely of the recall type, dealing with elements of chemical knowledge that the experienced teachers felt might reasonably be expected to have been presented to beginning chemistry students at some time in their prior schooling. From trial forms administered in September, 1962, a final form consisting of forty multiple choice items was produced, following essentially the same procedures as were used in developing the unit tests.

Results of Field Testing

During the 1963–64 school year, the pretest and unit tests were used by all of the participating teachers in their regular classes. To insure a usual level of motivation on the part of pupils taking the tests, the unit tests were used as regular, periodic tests of normal influence in determining grades. From a beginning population of about 1200 pupils enrolled in chemistry, complete data—including scores on the Chemistry Regents Examination—were obtained for 801, so that results obtained from these data should probably be considered as applicable only to rather healthy pupils (i.e., who are not absent from school for one or more tests) who do not move from one school district to another, and who actually finish the chemistry course.

Table II presents the intercorrelations among the unit tests and the chemistry pretest. Ranging from 0.45 to 0.78, with a median value of 0.69, it is considered that the tests are about as independent as the common thread of chemistry running through them would permit. Among the six unit tests, the range is 0.61 to 0.78, reflecting their greater similarity to one another than to the practically exclusively recall-type items of the pretest; the correlations of pretest scores with those on the unit tests, as might be expected, are the lowest in the matrix.

Since approximately half of the variance in scores for most of the pairs of unit tests

TABLE II
Intercorrelations among tests, $N = 801$

Test	2	3	4	5	6	7
1. Chemistry pretest	0.55	0.59	0.53	0.45	0.51	0.51
2. Periodic table test	—	0.71	0.72	0.61	0.69	0.70
3. Water, solutions, and ionization test		—	0.76	0.65	0.69	0.74
4. The halogen elements test			—	0.69	0.73	0.78
5. The sulfur–nitrogen test				—	0.72	0.68
6. Organic chemistry test					—	0.75
7. Principles of chemical reactions test						—

appeared to be shared variance, the prospects for efficient prediction of Regents Examination scores through multivariate analyses appeared favorable. Multiple regression analyses following the June, 1964, Regents Examination proved this to be the case, at least so far as the employment of two predictor variables was concerned; addition of predictors beyond two, however, did not yield correlations of sufficiently increased magnitude to justify the additional computational complexity. Multiple correlation coefficients ranged from 0.75 with the chemistry pretest and periodic table test as predicting instruments to 0.84, using the unit tests covering "principles of organic chemistry" and "principles of chemical reactions."

Whatever opinion one may hold regarding the statewide examining system in New York (and opinions are diverse), there is no denying the fact that performance of pupils on Regents Examinations *is* of considerable importance to teachers, counselors, parents, and the pupils themselves. For this reason, it is of local import to be able to make predictions regarding Regents scores, hopefully in the case of undesirable predictions, so that special steps may be taken to alter the predicted outcome. To facilitate the prediction process, the consultants prepared a series of nomographs depicting the two-predictor regression equations in graphic form,* and example of which is shown in Figure 1. Use of the nomographs by teachers and counselors is readily apparent, but it may well be, also, that they can be of direct help to students by enabling them to relate their individual progress to date to the Regents scores earned in the past by students with similar progress characteristics. The fact that the June, 1964 Chemistry Regents Examination was generally conceded to be of greater-than-usual difficulty should be advantageous in this regard.

Following the completion of all unit testing, four scores were derived for each student who had taken all six tests, representing the total number of items answered correctly on all tests for each of the four item categories. The intercorrelations among these scores are presented in Table III and will immediately be noted to be rather substantial for scores that presume to represent quite distinct intellectual competencies.

Considering the difficulty that was encountered in differentiating among items calling for abilities other than recall, it would not have been surprising to find high correlations among the three other categories. But the correlations between recall-type items and the other categories are equally substantial, indicating that a common factor is influential in determining performance on all four types of items. It may be, of course,

TABLE III
Intercorrelations among Item Types, $N = 801$

Item type	2	3	4
1. Recall	0.87	0.84	0.81
2. Comprehension	—	0.84	0.80
3. Application		—	0.82
4. Analysis, synthesis, and evaluation			—

* The nomographs were produced by a CalComp 570 digital plotter (California Computer Products) from data prepared by an IBM 7074 System.

that this factor is nothing more than *recall* (or memory), itself, which is an essential component of adequate functioning in any cognitive area. Other possible explanatory constructs would include general scholastic aptitude (or specific chemistry aptitude) that largely determines progress in all cognitive areas, or some combination of these.

Whatever the explanation, it is apparent to the individuals who have worked on this project that the practical application of the *Taxonomy of Educational Objectives*, in such a way as to produce meaningful results, is no easy task. With the data available to this investigation, no further conclusive explanation of the high correlations can be offered, but the finding does raise a definite question regarding the utility of the *Taxonomy* in test construction, particularly where the individual teacher depends upon his lone efforts for his own test.

Conclusions

As a means of dealing with the problems of developing achievement tests directed toward specific local requirements, the cooperative approach may certainly be recommended. Aside from the requirement for statistical demonstrations of test adequacy, the acid test of an achievement examination is teacher opinion, frequently influenced by pupil opinion, and in this respect, the tests developed as described here have passed. The tests are now being used as an integral part of the chemistry courses in the schools which participated in the study; the teachers like them, and it really is of little import that a large measure of this liking may be the result of their participation in the process of test development. The in service training value of the experience cannot be evaluated objectively, but it almost certainly is considerable.

On the other side of the coin, there is some question in the minds of the writers regarding the optimal size for a group effort such as this. It is probably not necessary to have as large a group as the one that worked on this project, and increased numbers multiply the problems of communication, meetings, unity of objectives, etc. Similarly, with regard to technical consultants to test development projects, there would probably be some advantage to having them affiliated with the same university, and it should not usually be necessary to have four of them.

Finally, note should be taken of the symbiotic properties of a project such as this one. Except in the largest school districts, technical specialists in test development are not available within the school structure; nor does one profess for long in the field of education in a university without feeling the need for more recent and intimate contact with the everyday problems and activities of public schools. Cooperative projects like this one can be mutually satisfying in these respects, and regardless of the ultimate value of being able to predict achievement (Regents or otherwise), if such predictions are going to be made, then they should be as efficient and accurate as possible, thus, if not benefiting the pupils, at least harming them less.

The project reported here was supported by the Office of Education Research, New York State Education Department, and the fourteen cooperating school districts.

References

1. Thorndike, R. L., and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education*, Wiley, New York, 1961, p. 289.
2. Ahmann, J. S., and M. D. Glock, *Evaluating Pupil Growth*, Allyn and Bacon, Boston, 1959, p. 350.
3. Nunnally, J. C., *Tests and Measurements*, McGraw-Hill, New York, 1959, pp. 270f.
4. Degering, E. F., et al., *Cooperative Objective Unit Tests in Organic Chemistry*, E. F. Degering, Natick, Mass., 1950.
5. Ahmann, J. S., and M. D. Glock, *Evaluating Pupil Growth*, Allyn and Bacon, Boston, 1959, p. 187.
6. Bloom, B. S., Ed., *Taxonomy of Educational Objectives, Part I: The Cognitive Domain*, David McKay, New York, 1956.
7. Dressell, P. L., and D. H. Nelson, *Questions*

and Problems in Science, Folio 1, Educational Testing Service, Princeton, 1956.

8. Gullikson, H., *Theory of Mental Tests*, Wiley, New York, 1950.

9. Winter, S. S., S. D. Farr, J. J. Monteau, and J. A. Schmitt, *A Multi-district Study of Achievement in New York State Regents Chemistry When Taught in Conventional and Large-group Classes*, The Educational Research Center, State University of New York, Buffalo, New York, 1965.

JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 4, PP. 92-94 (1966)

Hiding behind Course Titles*

FRANK X. SUTMAN
Temple University, Philadelphia, Pennsylvania

According to Brandwein,[1] among the major teaching skills which are rated excellent by their principals and colleagues are those skills useful in confronting students with interesting objects and events, in questioning them, and in leading discussion to sharpen their ability for designing a critical investigation. This description not only comes close to indicating the major desirable characteristics of science teachers today but also for tomorrow.

Introduction

In one's early days of teaching science, the opinion is usually held that brilliance coupled with knowledge of subject matter, is the only prerequisite to becoming an effective science teacher, a teacher with the characteristics just described. Experience and research indicate otherwise.[2,3]

What to some seems to be an easy solution to ineffective science teaching, truly, is not so easy. Today, however, we do have some results of reputable research to refer to for answers. I suggest we continue to pursue this research and begin to use the findings in developing a pattern for undergraduate experiences in the training of future secondary school science teachers. We should not continue to make the mistake of believing the answer to developing effective teachers lies solely in a list of prescribed courses!

Instead, it seems reasonable in years to come to establish a means of selecting from the population those people who we can predict with reasonable certainty will become effective science teachers, and then outline the experiences (not course titles) that will develop to the fullest possible extent in these people behaviors so essential to a productive, successful, and happy science teaching career.

After this, perhaps we might spell out the major environmental factors of the undergraduate schools and the first few years of teaching that will nurture and continue to develop good science teaching. And at the same time it might be well to prepare an alternate list of factors that destroy the will and the desire to mature in the complex art and science of science teaching.

Except for science content and the kind of practical information usually obtained in the undergraduate science methods course, our problem appears to be little different than that of many groups attempting to develop effective undergraduate training programs for teachers, since research findings indicate that superior qualities in teachers are common to many academic fields.[4]

Role of Content Courses

If, as already implied, the behavior patterns we seek can be developed through appropriate experiences in many sequences of content courses (and perhaps this is true as well for professional courses), we then should leave to the colleges the responsibility of deciding the nature and order of courses to be included in the undergraduate science

* Excerpted from an introductory paper presented at the Eastern Region Association for the Education of Teachers in Science meeting, Keene, New Hampshire, May 8, 1965.